Report:

Since I am a beginner I focused on completing the project without errors rather than competiting for best accuracy.

Methodology:

1. Import necessary libraries. I'll be using the lightGBM since it is relatively new and i read articles about it performing better than XGboost so i thought i would give it a try.

2. Import and observe data. Merge transaction set and identity set of both train and test dataset.

3. The RAM usage exceeds 16 GB if data is not handled properly. So, it is crucial to save memory wherever possible.

I converted floats into integers in columns wherever possible.

After merging, splitting etc operations, old dataframes were deleted.

4. There are a large number of missing values. First i calculated the columns that have missing values higher than 80% and dropped those columns in both test and train dataset. There were still missing values which i filled with the mode of the column

5. For preprocessing first i separated the object and numerical columns then implemented bundle preprocessing for numerical and categorical data using **simpleimputer** on numerical data and **simpleimputer + onehotencoder** on categorical data with pipeline

6. Finally the LightBGM model was called and hyperparameters were tuned via trail and error